学校编码：10384　　分类号___密级

学号：15420110153794　　UDC____

厦门大学

博 士 学 位 论 文

函数型数据挖掘的统计分类方法研究

Research on Statistical Classification Methods of Functional Data Mining

王 德 青

指导教师姓名： 朱 建 平 教 授
专 业 名 称： 统 计 学
论文提交日期： 2014 年3 月
论文答辩时间： 2014 年5 月
学位授予日期：

答辩委员会主席：_____
评 阅 人：_____

2014 年5月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名）:

2014 年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2014 年　　月　　日

摘　要

大数据时代已经来临成为社会各界的普遍共识，传统的数据分析技术在大数据时代的信息数据挖掘中面临诸多局限性，如何提出、修正和拓展适用于大数据的数据挖掘技术给现代统计学发展带来了机遇、挑战和紧迫感。函数型数据分析是一种研究如何从无穷维、不规则观测数据中挖掘内在信息知识的理论和方法，其核心思想是将观测数据视作随机过程的离散实现，从整体的函数视角对数据的变化模式进行分析；不仅放松了数据采集的结构约束和分布假设，而且可以对数据的深层次动态变化模式进行挖掘。分类是人们认识世界基本且重要的方法，基于函数视角的分类技术既能够挖掘传统的结构化数据信息，也能够探索非结构化数据的类别规律，对于丰富大数据时代的信息挖掘技术具有重要意义。本文针对函数型数据分类的特殊性展开讨论，从方法思想、理论模型和算法实践三个层面对函数型数据挖掘的统计分类技术进行研究。主要研究内容如下：

1. 界定函数型数据的概念与内涵、分析思想和统计特征。首先，系统分析函数型数据与传统数据的区别与联系，明确将传统统计分析技术直接推广至函数型情形的关键问题；其次，对比函数型主成分分析与经典主成分分析的异同，并案例展示函数型数据分析的核心优势，为后续章节中函数型数据的分类方法研究奠定理论基础。

2. 基于两步法的函数型聚类分析。首先，在假定观测数据能够用最优基函数展开的前提下，对函数化系数变量进行主成分降维，并将主成分距离按其重要程度自适应赋权加总作为函数之间的相似性测度，从离散角度对两步法的函数型聚类分析进行拓展；其次，针对最优基函数选择的主观任意性，在随机过程 Karhunen-Loève 展开基础上构建自适应权重的分类统计量，从连续角度对函数型自适应权重聚类分析进一步拓展；最后，实证分析验证了新方法的必要性、优良性和普遍适用性。

3. 基于曲线形状特征的函数型聚类分析。首先，剖析基于函数曲线形状特征分类的必要性，对比现有函数型非参数聚类分析的原理、优势与局限性。进一步，基于函数多角度的形状相似性提出一种基于曲线综合形状的半参数聚类分析。实证检验显示，新方法能够从动态角度深层次挖掘函数的类别规律，分类结

果更具稳定性。

4. 基于随机模型的函数型参数聚类分析。首先，剖析函数型迭代更新聚类分析的理论思想，明确初始类别中心确定稳健与否的重要性。在此基础上，应用自适应赋权主成分聚类分析对函数型迭代更新聚类模型的稳健性进行修正。统计模拟显示，修正的函数型自适应迭代更新聚类模型分类正确率明显提高。

5. 函数型数据的有监督分类方法。首先，在不改变 “距离”关系一一映射的基础上，将传统的多元 Fisher 判别分析推广至无穷维的函数空间；其次，对比不同窗宽选择和相似性测度下非参数判别分类正确率的差异，明确窗宽选择和相似性构建对函数型非参数判别模型的重要性；最后，针对高维线性判别分析处理函数型分类时存在的过拟合问题，在最优得分回归与线性判别分析等价的基础上，引入一种函数型惩罚判别分析，并对新模型的有效性进行了实证检验。

6. 多指标面板数据聚类分析及函数化角度的拓展。首先，系统总结面板数据聚类分析的原理、算法及存在的缺陷，在定义多指标面板数据静态和动态相似性测度的基础上，提出一种客观融合多角度相似性的多指标面板数据聚类分析；其次，将本文拓展的面板数据聚类模型应用于中国区域创新能力层级划分，分类结果的显著性检验显示，新方法能够科学划分区域之间的创新差异，依此制定的创新激励政策更具针对性。最后，基于连续的函数化角度介绍了多指标面板数据参数聚类分析的思想和具体步骤，并应用统计模拟技术验证基于随机模型多指标面板数据函数化聚类分析的有效性。

关键词：函数型数据挖掘；聚类分析；判别分析；自适应权重；面板数据

ABSTRACT

It is a general consensus on the advent of big data all over the world, and traditional data analysis techniques faces many limitations in big data mining, so presentation, revision and extension of mining technologies applicable to big data have brought opportunities, challenges and the urgency to the development for the development of modern statistics. Functional data analysis (FDA) is statistical theory and method aims to mining information and knowledge from the infinite dimensional, irregular data. Its core ideology is treating discrete observed data as discrete realizations of stochastic processes, and analyzing data's dynamic pattern from the overall perspective of function. For its merits, FDA not only relaxes structure constraints and distribution hypothesis of data acquisition, but also can mining dataset's deeper information dynamically. Classification is one of the basic and important methods, as for functional classification, it can mine information from traditional structural data. Besides, it also can explore the categories of unstructured data, so functional classification possesses an important status among mining technologies of big data. Aimed at the particularity of functional data classification, this paper analyses its classification ideology models and algorithms. The main research contents are as follows:

1. Defines FDA's concept connotation, analysis ideology and statistical features. First, Chapter 2 systematic analyzing the difference and connection between functional data with traditional data, and clarifying the key problems of generalizing traditional statistical methods to function data directly. Second, upon comparing the similarities and differences of traditional principal component analysis and functional principal components, Chapter 1 display case core advantage of functional data analysis, which lays a foundation for the study of functional classifications in the following chapters.

2. Functional clustering analysis based on two stages. First, under the assumption of observation data can be expanded by the basis function, Chapter 3 reduces

dimensions of functional coefficient by principal component analysis, and defines similarity measure between functions by weighting principal components adaptively, the above works give an extension of functional clustering based on two stages. Second, in order to solve the subjective arbitrary selection of basis function, we defines an adaptive weighting classification statistics by Karhunen-Loève of stochastic process, which gives deeper extension of adaptive weighting functional clustering from continuous perspective. Finally, the empirical analysis verifies the necessity excellence and universally application of the new method.

3. Functional clustering analysis based on curves' shape. First, this paper gives the necessity of classification based on curves' shape, and comprehensively compares the principle advantages and limitations of functional nonparametric clustering. Furthermore, we present a semi-parametric functional clustering algorithm based on functions' multi-perspective comprehensive shape similarity. The empirical test shows that the new methods can deeply mining category of functions from the dynamic perspective, and classification results more robust.

4. Model-based functional parametric clustering analysis. First, in Chapter 5 we give a comprehensive analysis on ideology of iterative-updated functional clustering analysis, and clarify the importance of selecting a robust initial cluster centers. Upon the above conclusion, we revise the robustness of iterative-updated functional clustering analysis combining adaptively weighting principal component clustering analysis. Statistical simulation shows the revised adaptively iterative-updated functional clustering model can promote classification correct rate significantly.

5. Functional supervised classification. First, without changing the "distance" upon the one-one mapping, we extend traditional Fisher discriminant analysis to infinite-dimension functional spaces. Second, we compare the classification correct rate of different bandwidths and similarity measures, , and clarify that bandwidth selection with similarity measure possessing an important status in clear and of function type non importance parameters of discriminant model in constructing functional nonparametric discriminant model. Finally, in order to avoid over-fitting problem when applying high dimension linear discriminant to processing functional

classification, we introduce a functional penalized discriminant analysis upon the equivalence of optimal score regression and linear discriminant analysis. Empirical test validates the effectiveness of new models.

6. Clustering analysis of multivariate panel data and its functional extension. First, we present a systematic summary on principle, algorithms and defects of panel data clustering analysis. Upon defining static and dynamic similarities of panel data, we put forward a multi-variant panel data clustering analysis by objectively combining multi similarities. Second, we apply the extended panel data clustering model to classify China's regional innovation ability. Significant test of the classification result shows that the new method can scientifically divide innovation difference among different regions, and innovation incentive policies made by the classification are more targeted. At last, we present the ideology and concrete steps of functional parametric clustering for multi-variant panel data from continuous perspective, and verify the effectiveness of model-based functional parametric clustering for multi-variant panel data.

.

Keywords: Functional Data Mining; Clustering Analysis; Discriminant Analysis; Adaptive Weighting； Panel Data

目 录

摘　要.....I

ABSTRACT.....III

第一章　绪 论.....1

1.1 选题背景与研究意义.....1

1.2 国内外相关研究现状.....3

1.3 研究内容与结构安排.....13

第二章　函数型数据分析.....19

2.1 函数型数据的概念与内涵.....19

2.2 函数型数据的统计描述.....22

2.3 离散数据的函数化处理.....24

2.4 函数型主成分分析.....27

2.5 函数型数据分析的案例应用.....29

2.6 本章小结.....32

第三章　基于自适应权重的两步法聚类分析.....33

3.1 问题的提出.....33

3.2 离散角度的函数型聚类分析评述.....34

3.3 自适应赋权主成分聚类分析.....36

3.4 新方法优良性的实证检验.....38

3.5 两步法聚类分析的连续角度拓展.....41

3.6 本章小结.....44

第四章　基于曲线形状特征的函数型聚类分析.....46

4.1 基于函数型秩相关的曲线形状聚类分析.....46

4.2 基于极值点属性特征的函数型聚类分析.....48

4.3 考虑时间因素的极值点属性聚类分析.....52

4.4 基于函数综合形状特征的聚类分析.....54

4.5 不同形状聚类模型的分类效果对比 58
4.6 本章小结 65
第五章 基于随机模型的函数型参数聚类分析 66
5.1 自适应权重迭代更新聚类分析 66
5.2 函数型数据的概率密度 70
5.3 基于概率密度的参数聚类分析 72
5.4 模型有效性的检验与应用 74
5.5 本章小结 82
第六章 函数型数据的有监督分类方法 84
6.1 经典多元判别分析 84
6.2 基于两步法的函数型判别分析 86
6.3 函数型非参数判别分析 88
6.4 函数型惩罚判别分析 95
6.5 本章小结 99
第七章 面板数据的聚类分析及其函数化拓展 101
7.1 问题的提出 101
7.2 单指标面板数据聚类分析 101
7.3 多指标面板数据聚类分析 103
7.4 基于随机模型的面板数据函数化聚类分析 115
7.5 本章小结 121
第八章 文章总结与研究展望 123
8.1 全文总结 123
8.2 论文不足与研究展望 125
参考文献 127
论文发表与科研项目 138
致 谢 139

TABLE OF CONTENTS

Chinese Abstract ..... **I**

Abstract ..... **III**

Chapter 1 Introduction **1**

1.1 Research Background and Significance 1

1.2 Research Status at Home and Abroad 3

1.3 Contents and Framework 13

Chapter 2 Functional Data Analysis **18**

2.1 Concepts and Connotation of FD 19

2.2 Statistical Description of FD 22

2.3 Functionalization of Discrete Data 24

2.4 Functional Principal Component Analysis 27

2.5 Application Exhibition of FDA 29

2.6 Summary 32

Chapter 3 Functional Cluster Based on Adaptive Two-stage ... **33**

3.1 Problem Presentation 33

3.2 Reviews of Functional Cluster from Discrete Perspective 34

3.3 Adaptively Weighting PCA Cluster Analysis 36

3.4 Empirical Test of New Methods' Superiority 38

3.5 Continuous Extension of Two-stage Cluster Analysis 41

3.6 Summary 44

Chapter 4 Functional Cluster Based on Curves' Shape **46**

4.1 Shape Cluster Based on Functional Rank Correlation 46

4.2 Functional Cluster Based on Extreme Points 46

4.3 Functional Cluster Based on Extreme Points Involved Time 51

4.4 Cluster Based on Function's Comprehensive Shape 53

4.5 Clssification Comparison of Different Shape Cluster Models 57

4.6 Summary 64

Chapter 5 Model-based Parametric Cluster of FD 65

5.1 Adaptively Weighting Iteration Cluster 65

5.2 Probability Density for Random Functions 69

5.3 Parametric Cluster based on Probability Density 71

5.4 Test and Application of Models' Effevtiveness 73

5.5 Summary 81

Chapter 6 Supervised Classification of Functional Data 83

6.1 Classical Multivariant Discriminant Analysis 83

6.2 Functional Discriminant Analysis based on Two Stages 85

6.3 Functional Nonparametric Discriminant Analysis 87

6.4 Functional Penalized Discriminant Analysis 94

6.5 Summary 98

Chapter 7 Cluster of Panel Data and Its Functional Extension 99

7.1 Problem Presentation 99

7.2 Cluster of Univariate Panel Data 99

7.3 Cluster of Multivariate Panel Data 101

7.4 Functional Cluster of Panel Data based on Model 113

7.5 Summary 119

Chapter 8 Research Summary and Expectation 121

8.1 Summary of Dissertation 121

8.2 Limitations and Future Research Issues 123

References 125

Publications and Research Projects 135

Acknowledgements 136

第一章　绪 论

1.1 选题背景与研究意义

随着人类认识世界的不断深入、现代信息技术的迅猛发展以及数据存储能力的极大提高，自然科学和社会科学的许多领域不断涌现出大量形式各异复杂难辨的海量数据，如（超）高维数据、（超）高频数据、不等观测时点数据、非平衡数据等，标志着人类进入一个全新的时代——大数据时代。“数据爆炸”和“知识匮乏”是大数据时代的典型特征，一方面，数据采集技术的进步和存储成本的下降使得数据容量的起始单位由 GB 变为 ZB[①]；数据类型不仅有数字、符号等结构化数据，而且还有视频、图片等非结构化数据。另一方面，数据维度容量的庞大和类型结构的复杂使得传统的数据分析技术凸显诸多局限，甚至完全失效。因此，如何从浩瀚复杂的数据海洋中及时有效地挖掘出潜在的深层次信息，给现代统计学的发展带来了挑战、机遇和紧迫感。

大数据时代，数据的丰富性和多样性对高效的数据分析技术提出了更高的要求。传统数据分析技术的研究重点主要集中在时间序列数据（Time Series Data）、横截面数据(Cross-sectional Data)或者二者的综合—多指标面板数据(Multivariate Panel Data)，从线性模型到非线性模型、从低维空间到（超）高维空间、从等间隔观测的平衡数据到不等间隔观测的非平衡数据等，理论方法和实践应用的研究成果都是有针对性地处理某类特定的数据类型，诸多的理论假设条件导致模型应用的普适性较差。主要表现在：一是以线性结构为模型变量之间的主要形式，限制了复杂系统的非线性、非平稳等不规则变化描述，不能真实地反映系统运行的真实情况；二是过分依赖大量的经典假设，如变量的平稳性、独立性、数据等间隔观测等，一旦假设条件遭到破坏则模型应用的有效性急剧下降，依据模型结果得出政策建议的可靠性有待商榷；三是数据生成过程（Data Generate Process）的信息含量不足，依据有限离散样本数据建立模型的外延性预

[①]数据的计算单位为：1024EB=1ZB，1024PB=1EB，1024TB=1PB，1024GB=1TB

测受到极大限制。事实上，大多数数据生成过程是连续的动态过程，可以用随机过程的数学模型进行描述，而实际观测的数据往往是稀疏不一且带有噪音影响、离散的有序排列。如果将初始观测数据视作一般的静态数据，不加处理（如噪声消除等）直接使用传统的数据分析方法进行研究，则会造成重要信息量损失或模型估计失真等严重问题。针对传统数据分析技术的上述缺陷，一种全新的数据分析思路——函数型数据分析应运而生。

广义上讲，时间序列数据、横截面数据和（多指标）面板数据都是函数型数据的特殊形式，传统类型数据分析的思想、理论和方法都可以在函数型数据情形下得到推广和拓展。直观角度上，多指标描述的样本在特定时点上的取值构成截面数据，在样本容量保持不变的前提下，特定变量在时间维度纵向离散取值则为时间序列数据，时间序列数据在横向的多维角度等间隔扩充则构成面板数据（或普遍意义上的纵向数据），传统的计量方法在处理上述类型的数据已有成熟的理论和方法。但应该注意到，实际问题中的数据生成往往具有前后继承性和内在逻辑结构，对于任一研究对象在不同时间和空间的多次观测值，往往可以通过原始观测数据重构隐含在其中的本征函数。进一步地，基于连续函数视角进行数据分析能够挖掘出更多有价值的信息。具体地，与传统的数据分析方法相比，函数型数据分析的核心优势主要体现在：①以连续的函数为分析对象，能够从整体上把握数据的分布特征和变化规律；②依赖较少的假设条件和较弱的结构约束，挖掘模型的普遍适用性显著提高；③ 不要求所有观测对象在相同时点等间隔取值，数据采集和处理方式更具灵活性和方便性；④可以从高阶的导数函数（或微分方程）角度对系统的运行规律进行多角度的静态与动态分析，信息挖掘的深度明显提升；⑤不仅能够分析传统的结构化数据，也能够处理诸如音频、笔迹等非结构化数据。综合来看，基于函数角度的数据分析技术是大数据时代信息挖掘方法研究的一个新颖视角，函数的连续动态性给数据分析带来了挖掘角度的便利性和结论建议的针对性，但函数型数据的无穷维特征也给统计推断和算法实现带来了极大的挑战性。因此，发展适用于函数型数据的统计分析方法对于丰富大数据时代的数据挖掘技术具有重要的理论研究价值和广泛的应用前景。

目前，函数型数据分析的理论研究和应用研究正处于蓬勃发展阶段，面向函